Applied Physics 应用物理, 2014, 4, 101-106
Published Online June 2014 in Hans. http://www.hanspub.org/journal/app
http://dx.doi.org/10.12677/app.2014.46014

Hans汉斯

Link Prediction Based on Clustering Coefficient

Zixuan Huang, Chao Ma, Jinhui Xu, Jiangnan Huang

Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou
Email: 313260224@qq.com

Received: Apr. 28th, 2014; revised: May 26th, 2014; accepted: June 5th, 2014

Abstract

Link prediction in complex network aims at estimating the likelihood of the existence of links between nodes by the known network structure. Currently, most link prediction algorithms are similarity algorithms based on local information including the number of common neighbor nodes, degree of common neighbor nodes and the interactions between common neighbor nodes, and thus their applied range is limited. In this paper, we consider the interactions between adjacent nodes of a node and design a new algorithm based on clustering coefficient. We use this new algorithm in the experiments on real networks and simulative networks generated by pajek, and experimental results show that the algorithm is applicable to a wide range of problems and it has the high accuracy of prediction.

Keywords

Complex Network, Link Prediction, Clustering Coefficient

基于集聚系数的链路预测算法

黄子轩，马　超，徐瑾辉，黄江楠

广东外语外贸大学思科信息学院，广州
Email: 313260224@qq.com

收稿日期：2014年4月28日；修回日期：2014年5月26日；录用日期：2014年6月5日

摘　要

复杂网络中的链路预测是指基于已知的网络结构信息来预测网络中尚未链接的两个节点间产生连边的可能性。现有算法主要是基于局部信息的相似性算法，即针对共同邻居的数量、共同邻居的度值以及共同邻居之间的相互链接程度进行研究，应用范围有限。为此，针对一个节点的邻接点之间相互链接的程度，本文提出一种基于集聚系数的新算法。本文利用该算法对多种现实网络以及pajek生成的模拟网络进行了实验，实验结果表明，该算法适用范围广，链路预测准确率高。

关键词

复杂网络，链路预测，集聚系数

1. 引言

如何理解复杂网络的演化是当前研究的热点，而链路预测是网络演化中的一个基本问题。由于节点的属性信息难以获取，预测算法的准确率难以得到保证。同时，在现实中不少网络规模较大，增大了计算开销。由于这个原因，基于局部信息的相似性指标的链路预测算法受到了更大的关注[1]。早期的链路预测研究思想以基于共同邻居的个数为主，即：若两个节点之间存在更多的共同邻居，则该两个节点有更大的可能产生连边。而在共同邻居个数的基础上，根据两端节点度的影响从而逐渐发展出以下指标：Salton 指标、Jaccard 指标、Sorenson 指标、大度节点有利(HPI)指标、大度节点不利(HDI)指标、LHN-I 指标[2]。随后，一种以共同邻居的度值为基础的思路被提出，从而产生了 Adamic-adar (AA)指标和资源分配(RA)指标[3]。

随着研究的不断深入，很多学者注意到不能单纯地考虑共同邻居的个数或其度值，而是应该深入研究共同邻居之间的相互关系，即：若共同邻居之间存在更多的连边，则两端节点属于同一个社区的可能性更大，产生连边的可能性也会更大[4]-[6]。

通过对 10 种基于节点局部信息的相似性指标的研究，吕琳媛发现 CN 指标、Adamic-adar (AA)指标和资源分配(RA)指标算法[7]比其余算法有较好的准确率[3]。这三种指标分别定义如下：

(1) CN 指标：基于节点局部信息的最简单的相似性指标是共同邻居指标(common neighbors)。该指标定义为两个节点的共同邻居数量，即两个节点之间如果有更多的共同邻居，则更倾向于连接。对于节点 i，定义 i 的邻居集合为 $\Gamma(i)$，设节点 x 和节点 y 的相似性为 S_{xy}，即：

$$S_{xy} = \left|\Gamma(x) \cap \Gamma(y)\right| \tag{1}$$

(2) AA 指标：该指标考虑的是两个节点的共同邻居的度值信息，其思想为度值小的节点的贡献度大于度值大的节点。因此根据共同邻居节点的度值给每个节点赋予一个权重，该权重为节点度值得对数分之一：$\frac{1}{\lg k}$。对于节点 i，定义 i 的邻居集合为 $\Gamma(i)$，$k(i)$为节点度值。设节点 x 和节点 y 的相似性为 S_{xy}，即：

$$S_{xy} = \sum \frac{1}{\lg k(z)},\ k \in \Gamma(x) \cap \Gamma(y) \tag{2}$$

(3) RA 指标：该指标是从网络资源分配的角度提出的。考虑网络结构中尚未连接的两个节点 x 和 y，从 x 传递资源到 y。而在传递过程中，它们的共同邻居充当传递的媒介。假设每个共同邻居作为媒介时能

将一单位的资源平均分配到它的邻居，则节点 x 和节点 y 的相似度为 S_{xy} 定义为节点 y 能接收到从节点 x 开始传递的资源数。对于节点 i，定义 i 的邻居集合为 $\Gamma(i)$，$k(i)$为节点度值。设节点 x 和节点 y 的相似度为 S_{xy}，即：

$$S_{xy}=\sum\frac{1}{k(z)},\ k\in\Gamma(x)\cap\Gamma(y) \tag{3}$$

以上三种指标虽然准确率较高，但依然有着一定局限性。CN 算法将所有的共同邻居等同看待，仅利用共同邻居的数量作为节点对间相似度的评分函数，没有区分出不同的邻居对链接预测的影响是不一样的；AA 算法和 RA 算法虽然区分了每个不同的共同邻居对链接预测的不同的影响力，但是它们都只关注于共同邻居本身，而忽略了这些共同邻居之间的相互影响。

为了克服这些局限性，本文提出了基于集聚系数的链路预测方法。该算法所基于的假设是，节点的集聚系数越大，则尚未相连的邻接点间存在连边的概率越大。基于这个思想，本文提出一种基于集聚系数的新算法，该算法同时考虑了两端节点以及共同邻居的本身特征，能反应出节点间的相互作用关系。

2. 问题描述

定义 $G(V,E)$为一个无向网络，其中 V 为节点集合，E 为边集合。网络的总节点数为 N，边数为 M。该网络共有 $N(N-1)/2$ 个节点对，即全集 U。这样，链路预测问题可描述为：构建某种相似度指标，对没有连边的节点对(x,y)计算其相似度 S_{xy}，S_{xy} 越大，节点对(x,y)出现连边的概率越大。

为了衡量算法的精确度，我们将已知连边 E 分为训练集 E^T 和测试集 E^P 两部分，其中 E^P 为在 E 中随机抽取的 10%的边，剩下的边均匀分配至 E^T。另外，将属于 U 但不属于 E 的边定义为不存在的边。本文使用 AUC 指标来衡量算法的准确性。AUC (area under the receiver operating characteristic curve)定义为随机在测试集中选取的边的分数值比随机选择的一条不存在的边的分数值高的概率。即每次随机从测试集中选取一条边与随机选择的一条不存在的边进行比较，加分规则如下：若测试集的边的分数值大于不存在的边的分数值，则加 1 分；若两个分数值相等，则加 0.5 分；若测试集的边的分数值小于不存在的边的分数值，则加 0 分。重复以上步骤，独立地比较 n 次，定义 n_1 为测试集中边的分数值大于不存在的边的分数值的次数，定义 n_2 为两个分数值相等的次数，则定义 AUC 为：

$$\mathrm{AUC}=\frac{n_1+0.5n_2}{n} \tag{4}$$

3. 基于集聚系数的链路预测算法

本节提出一种新的基于局部信息相似性的链路预测算法，该算法综合考虑了两端节点以及共同邻居的本身特征，同时由于集聚系数的特性，即它能反应出节点间的相互关系，所以该算法有着更高的参考价值。

集聚系数(clustering coefficient)用来描述一个节点的邻接点之间相互连接的程度，将该系数运用在社交网络中，其意义为该用户的朋友之间也是朋友的概率[8]。设网络中一个节点 i，该节点的集聚系数 $C(i)$ 等于其邻接点之间连边的数量除以邻接点之间可以连出的最大边数。设 e_i 为节点 i 的邻接点之间连边的数量，$k(i)$为节点 i 的度值，则：

$$C(i)=\frac{2e_i}{k(i)(k(i)-1)} \tag{5}$$

本文认为，节点的集聚系数越大，其邻接点集结成团的概率越大，则尚未相连的邻接点间存在连边

的概率越大。因此，本文提出一种基于集聚系数的新算法，该算法在考虑共同邻居数量的同时，还考虑了节点的集聚系数，即对每个参与计算的节点进行赋权，算法中节点 x 和节点 y 的相似性定义为 Clustering Coefficient(CC)指标：

$$S_{xy} = \sum\left(1+C(i)\right),\ i \in \left|\Gamma(x) \cap \Gamma(y)\right| \tag{6}$$

CC 算法流程图如图 1 所示。

4. 实验结果

4.1. 数据集

为对本文所提算法进行测试，本文使用了 4 种现实网络数据：adjnoun 网络、football 网络、polbooks 网络、US power grid 网络以及 3 种由 pajek 生成的模拟小世界网络：SW1、SW2、SW3。它们的网络拓扑性质见表 1。其中，N 和 M 分别表示网络的节点数和边数，K 表示网络平均度，C 表示网络集聚系数，r 表示网络同配系数，e 表示网络效率。

4.2. 实验结果

将 CN、AA、RA、CC 在 7 个网络中进行实验，并比较准确率。在 AUC 测试正确率中，对原始数据集进行了随机抽取分成了训练集(含 90%的连边数)和测试集(含 10%的连边数)，随后进行了 10^6 次的随机抽样比较。

4.3. 实验分析

我们使用 AUC 评价指标，对三种经典方法和本文所提方法进行对比试验，实验结果如表 2 所示。结

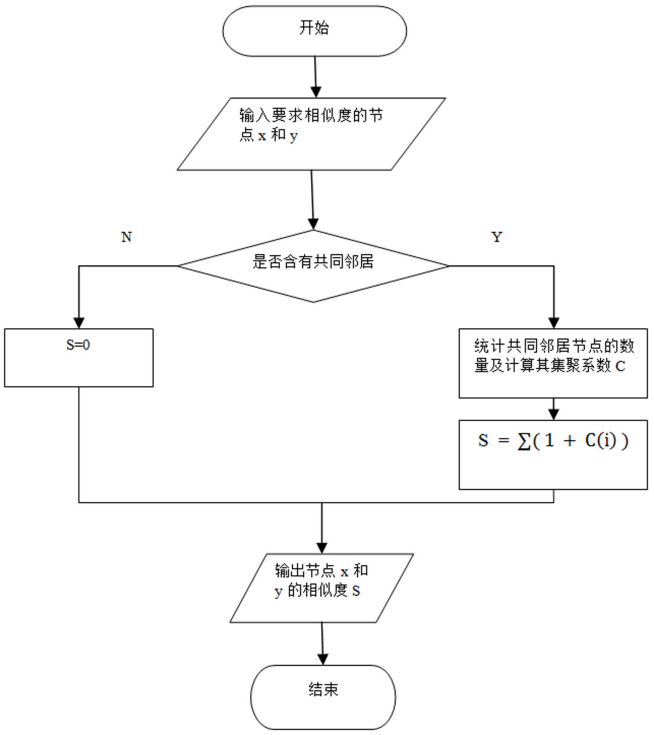

Figure 1. Flowchart of CC
图 1. CC 流程图

Table 1. Topological properties of seven networks
表 1. 7 个网络的拓扑性质

Nets	N	M	K	C	r	e
adjnoun	112	430	7.6786	0.2176	-0.1261	0.0692
football	115	613	10.6609	0.4032	0.1624	0.0935
polbooks	105	441	8.4000	0.4875	-0.1279	0.0808
US power grid	4941	6594	2.6691	0.0801	0.0035	5.4030e-004
SW1	1000	2000	4.0000	0.0573	0.0018	0.0040
SW2	1000	3000	6.0000	0.0756	0.0026	0.0060
SW3	1000	4000	8.0000	0.0982	0.0116	0.0080

Table 2. Experimental result (AUC) n = 1,000,000
表 2. 实验结果 (AUC) n = 1,000,000

Algorithm	adjnoun	football	polbooks	grid	SW1	SW2	SW3
CN	**0.6876**	0.8936	0.9096	**0.5868**	0.5740	0.6174	0.6657
AA	0.6777	**0.8945**	**0.9175**	0.5862	**0.5744**	**0.6178**	**0.6667**
RA	0.6766	0.8944	**0.9163**	0.5857	0.5739	**0.6176**	0.6659
CC	**0.6943**	**0.8971**	0.9159	**0.5873**	**0.5742**	0.6173	**0.6665**

果表明，本文所提出的 CC 指标在 7 种网络中都取得较好的准确度。该指标在 adjnoun 网络、football 网络和 US power grid 网络中有最高的正确率，在 SW1 网络和 SW3 网络有第二的正确率，在 polbooks 网络和 SW2 网络的正确率则与其他算法较为接近。从表 1 的网络拓扑信息中可以发现， 由于 US power grid 网络和三个模拟生成的网络均具有较低的网络平均度和平均集聚系数，因此局部信息相对较少，从而影响到指标的准确率。总体来看，CC 指标在 7 个网络中表现稳定，且在部分网络中表现突出。

5. 总结

传统的基于局部相似性指标的链路预测方法，主要单独考虑节点相似性或路径相似性因素。为了将二者更好地结合，进一步提高预测算法的准确率与稳定性，本文提出了基于集聚系数的链路预测算法。该法综合考虑了两端节点以及共同邻居的本身特征，能反应出节点间的相互作用关系。我们的实验结果表明：该方法具有准确率高、表现稳定等特点，对链路预测方法及应用有着重要的参考价值。

致　谢

本论文感谢广东外语外贸大学大学生创新创业训练计划项目的支持。

参考文献 (References)

[1] 吕琳媛, 陆君安, 张子柯, 闫小勇, 吴晔, 史定华, 周海平, 方锦清, 周涛 (2010) 复杂网络观察. *复杂系统与复杂性科学*, **2-3**, 173-186.

[2] Liben-Nowell, D. and Kleinberg, J. (2007) The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, **58**, 1019-1031.

[3] 吕琳媛 (2010) 复杂网络链路预测. *电子科技大学学报*, **5**, 651-661.

[4] 东昱晓, 柯庆, 吴斌 (2011) 基于节点相似性的链接预测. *计算机科学*, **7**, 162-164.

[5] 殷涵 (2012) 社会网络的链接预测. 硕士论文, 吉林大学, 吉林.

[6] 李淑玲 (2012) 基于相似性的链接预测方法研究. 硕士论文, 哈尔滨工程大学, 哈尔滨.

[7] Zhou, T., Lü, L. and Zhang Y.-C. (2009) Predicting missing links via local information. *European Physical Journal B*, **71**, 623-630.

[8] Feng, X., J.C. Zhao, J.C. and Xu, K. (2012) Link prediction in complex networks: A clustering perspective. *European Physical Journal B*, **85**, 3.